Introduction:

The COVID19 disease associated with the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection resulted in variable and heterogeneous symptoms. [1] This heterogeneity resulted in poor predictions from machine learning models when applied to general population to predict the risk of being admitted, and prognosis of disease for patients with COVID-19. The heterogeneity in the clinical phenotypes can be identified using clustering algorithms in machine learning models that is an unsupervised method and does not need any hypothesis of patients' severity of disease by the clinician resulting in a data-driven approach without any bias. Several cluster analyses of long COVID based on various phenotypes like respiratory cluster, fatigue cluster etc. have been reported, and explored for association with background factors and quality of life. [2], [3], [4] This study aims to perform a cluster analysis of clinical data for patients with COVID-19 in North Carolina with labs and vitals available within 14 days since the start of hospitalization. It considered data collected at the time of the hospital visit only, and an average across 14 days since the hospital visit. The clustering results were compared to determine the best method of data collection and number of clusters for grouping COVID-19 patients to evaluate the heterogeneity in phenotypes.

Method:

The data comprised of patient id, demographic information like gender, race/ethnicity, age at the time of hospitalization, and the measurements for 22 vitals and laboratory tests taken within 14 days since the start of visit. One thing to note is that not all patients had all the 22 labs and vitals done on all the 14 days leading to missing data in the dataset. Also, some patients had multiple readings for a type of measurement on a single day. These values were averaged over the day. Subsequently all the readings were consolidated by days (0 to 14 days) since the start of the hospital visit for each patient for all of the 22 labs and vitals.

The distribution of demographics for the patients is provided in Table 1. The details on the classification as mild, moderate and severe COVID are provided in earlier assignment applied concept paper1. For this study we are interested in understanding patterns in clinical data by employing unsupervised clustering method, KMeans implemented in the Python library scikit-learn. [5] Since this is an unsupervised learning the categorization as mild, moderate and severe were not provided to the KMeans algorithm during the clustering.

Table 1 Summarized demographic (sex, age, and race) information of 1763 patients in the sample categorized as Mild COVID, Moderate COVID and Severe COVID. No matching concept implies unknown race.

		Mild COVID n=1,003	Moderate COVID n=541	Severe COVID n=219
sex	female	518(52%)	299(55%)	92(42%)
	male	485(48%)	242(45%)	127(58%)
age	<18	186(19%)	35(6%)	2(1%)
	18-45	403(40%)	159(29%)	14(6%)
	46-65	234(23%)	204(38%)	62(28%)
	>65	180(18%)	143(26%)	141(64%)
race	asian	57(6%)	40(7%)	17(8%)
	black or african american	113(11%)	38(7%)	12(5%)
	no matching concept	7(1%)	6(1%)	NaN
	white	826(82%)	457(84%)	190(87%)

In order to prepare the data for clustering it was further summarized at patient level for all the 22 vitals and labs. This collapsed dataset can be created in different ways, and the ones tried for this study are - Case1: the vitals and labs at the day of the visit only, Case2: average the measurements over first 7 days, Case3: average the last 7 days, and Case4: average across the 14 days. During the dataset preparation for clustering, those patients without at least one of the 22 vitals and labs, in the summarized dataset, were categorized as NotAssigned since these patients were not included in the dataset provided to KMeans for clustering. Among the 1763 patients only 5 were categorized as NotAssigned in Case1 as most of the patients had one of the 22 vitals and labs taken on the day of the hospital visit. However, we see less variation in clinical test results among the patients on the day of the visit, refer ACP1 (Applied Concept Paper1) hence Case1 might not result in well separated clusters. In Case4 we retained all of the 1763 patients. Since the clinical results for the mild Covid patients are scarce, we would drop less of these patients by averaging across 14 days. The Case2 and Case3 were decided based the longitudinal study of clinical data by days since hospital visit (refer ACP1). Within 7 days the mild covid patient's clinical results moved towards normal values hence this period could be good for differentiating mild from the medium and severe patients. And after 7 days we see an improvement in clinical data of medium Covid patients that would help in differentiating medium from severe patients. In Case2, only 3 patients were dropped from the dataset and were marked as NotAssigned, but in Case3 (last 7 days) this number was significantly high (991 patients) and most of those dropped due to lack of data were mild COVID patients.

The data for clustering was preprocessed by imputing the missing values with the mean values of the respective features, followed by normalization employing StandardScaler in the Python library scikit-learn. [5] Since all 22 features might not be relevant, it was reduced to a set of relevant principal components that captured 80% of the variation in the data by performing linear dimensionality reduction using the PCA() of scikit-learn. See Figure 1 as an example for PCA analysis of Case4 where 13 principal components explained 80% of the data, hence the dimension of the dataset in this case was reduced from 1763 x 22 to 1763 x 13. The optimal number of clusters was determined by applying elbow method and silhouette analysis for clustering method evaluation. The identification of the bend(elbow) in the plot of the sum of squared distance against the number of clusters was visually challenging so it was determined by using

the KneeLocator from the Kneed library in Python. In Silhouette analysis the average silhouette coefficients were computed by varying the number of clusters from 2 to 30, and those that resulted in high value of coefficients were selected for clustering. After performing clustering on the dataset the differences between the clusters were shown by plotting the distribution of demographics and frequency of clinical measurements within the clusters.

Result:

In this report I will discuss clustering from Case1 and Case4 dataset. The Case1 dataset was created by selecting measurements on the day of the hospital visit, and the Case4 was created by averaging across 14 days since the start of visit. These cases were selected because few patients out of 1763 patients were dropped due to lack of data. For instance, 5 patients got dropped in Case1 and none in Case4 compared to 991 in Case3 (average last 7 days) and 3 in Case2 (average first 7 days).

The features were reduced by selecting those principal components that contributed to 80% of the variance in the data, see Figure 1. The 22 clinical measurement types were reduced to 16 features in Case1, and 13 in Case4. The number of clusters for a clustering model was selected by two cluster evaluation methods - the elbow method and the silhouette method, see Figure 2 and Figure 3 for Case1 and Case4 respectively. In the elbow method the location of the elbow was not completely obvious, so the Kneedle algorithm was applied to locate the inflection point in the line plot between total within-cluster SSE (sum of squared error) vs number of clusters. SSE is a measure of how close the points are from its cluster centroid. The optimized number of clusters for Case1 and Case4 are 10 and 7 respectively (see left figures in Figure 2 and Figure 3).

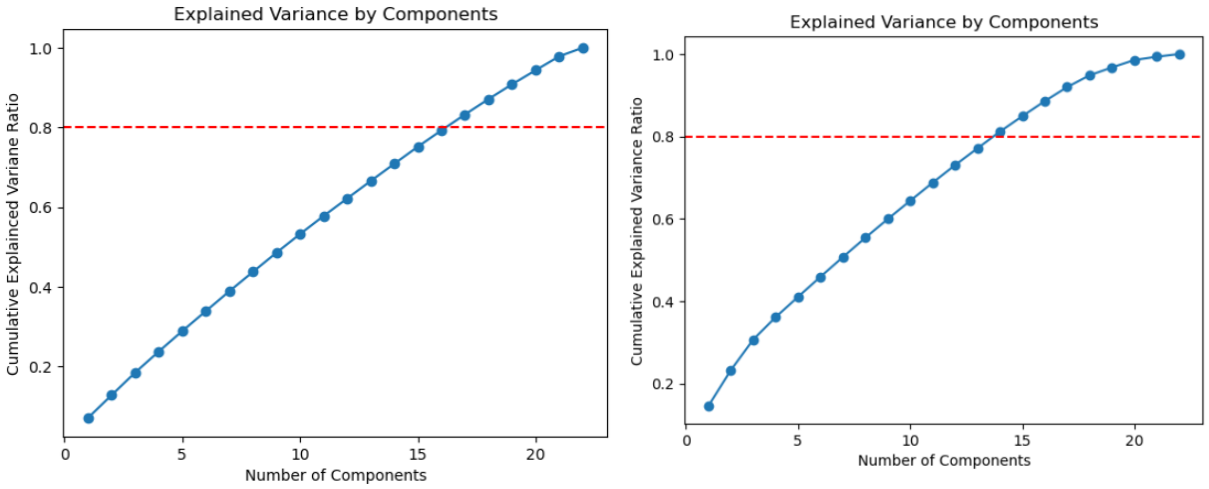

Figure 1 Left: the principal component analysis (PCA) on the dataset for Case1 (first day only), and right: PCA on the dataset for Case4 (average across 14 days). The plot of cumulative explained variance ratio of the data against the number of principal components. The red dashed line indicates the 80% of the explained variance of the data.

The second method, Silhouette method to determine the number of clusters involved plotting average silhouette coefficient, a measure of how separate the clusters are from each other, vs number of clusters. The ideal clustering would result in a value close to 1. In both cases Case1 and Case4, the optimal number of clusters are 2. However, in the Case4 we also see that number of clusters as 4 is the second-best choice (see right figure in Figure 3). Sometimes domain knowledge is considered when determining the ideal number of clusters, and based on domain knowledge and longitudinal study of clinical data (ACP1) three

clusters (mild, moderate and severe) were also considered. So, the choices for the number of clusters are 2 and 3 for Case1 and 2,3 and 4 for Case4. These seemed more reasonable compared to 10 and 7 that we obtained from the elbow method, consequently any differences in characteristics like demographics and clinical data were determined for 2 and 3 clusters for Case1 (see Figure 4 and Figure 5)and 2,3 and 4 clusters for Case4 (see Figure 6, Figure 7 and Figure 8). The demographics of the clusters provided information on any differences in age, distribution of race/ethnicity and distribution of gender. Also, we determined the frequency (as percentage) of each of the 22 labs and vitals in the cluster that would provide information on the severity of COVID.

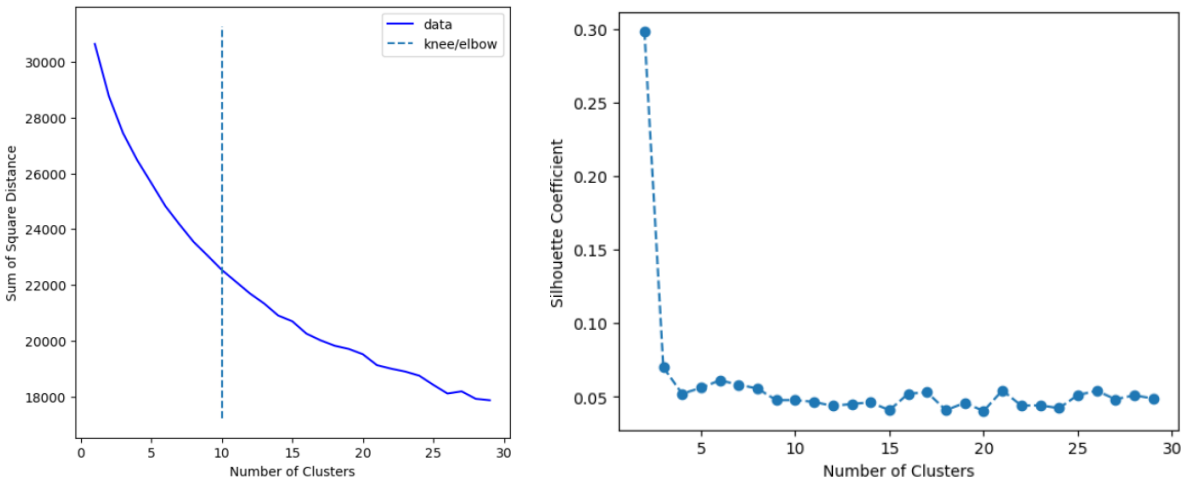

Figure 2 Left: The total within-cluster sum of square distances against the number of clusters and the location of elbow shown in blue dashed line, and Right: average silhouette coefficients by the number of clusters for Case1 where only the clinical data at the time of the hospital visit was considered.

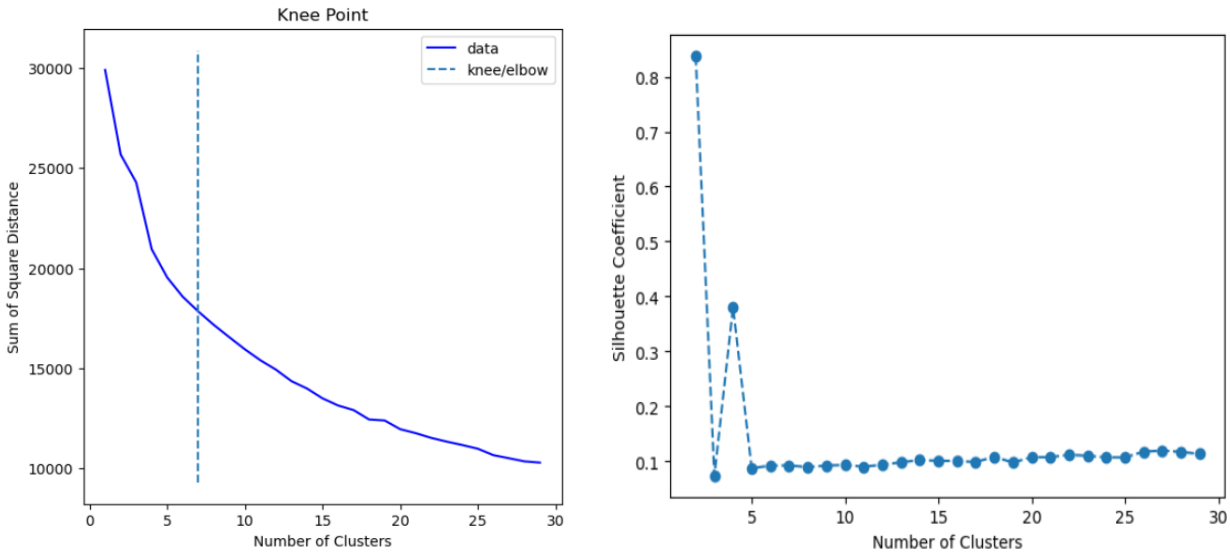

Figure 3 Left: the total within-cluster sum of square distances against the number of clusters and the location of elbow shown in blue dashed line. Right: average silhouette coefficients by the number of clusters for Case4 (average across the 14 days).

In Case1, the two clusters differed slightly in age and gender as Cluster1 was slightly older and predominantly male than Cluster0 (see Figure 4). Approx 5 younger patients (approx. 41 years) got dropped as they didn't have any of the 22 clinical data reported at the start of hospitalization. The two clusters have very similar frequency of clinical data. Note there are 23 times more patients in Cluster0 than in Cluster1. In case of three clusters, the Cluster0 splits and creates another Cluster2, further reducing the differences among the clusters, see Figure 5.

	Overall	0	1	NotAssigned
n	1763	1686	72	5
Age	48.4 ± 25.0	48.0 ± 25.0	58.2 ± 21.8	40.6 ± 32.9
Female	909 (51.6%)	874 (51.8%)	33 (45.8%)	2 (40.0%)
White	1473 (83.6%)	1414 (83.9%)	56 (77.8%)	3 (60.0%)
Black	163 (9.2%)	153 (9.1%)	8 (11.1%)	2 (40.0%)
Asian	114 (6.5%)	108 (6.4%)	6 (8.3%)	0 (0.0%)
Other	13 (0.7%)	11 (0.7%)	2 (2.8%)	0 (0.0%)

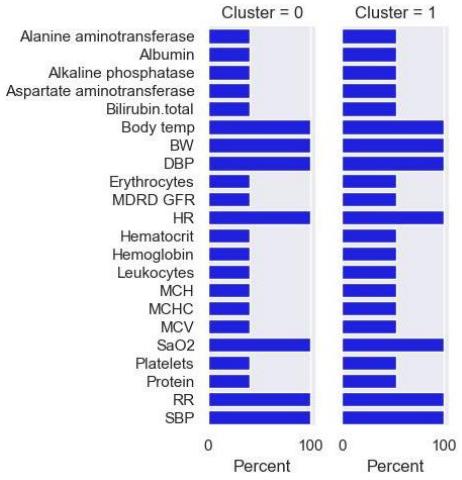

Figure 4 Left: The distribution of demographics within the two clusters and the overall patients in Case1 (0 day of hospital visit only). Those patients labeled as NonAssigned were not considered for clustering due to absence of any clinical data at the time of the visit. Right: The bar graph depicts the frequency of each measurement within the clusters.

	Overall	0	1	2	NotAssigned
n	1763	884	72	802	5
Age	48.4 ± 25.0	49.0 ± 24.2	58.2 ± 21.8	46.9 ± 25.7	40.6 ± 32.9
Female	909 (51.6%)	440 (49.8%)	33 (45.8%)	434 (54.1%)	2 (40.0%)
White	1473 (83.6%)	731 (82.7%)	56 (77.8%)	683 (85.2%)	3 (60.0%)
Black	163 (9.2%)	95 (10.7%)	8 (11.1%)	58 (7.2%)	2 (40.0%)
Asian	114 (6.5%)	50 (5.7%)	6 (8.3%)	58 (7.2%)	0 (0.0%)
Other	13 (0.7%)	8 (0.9%)	2 (2.8%)	3 (0.4%)	0 (0.0%)

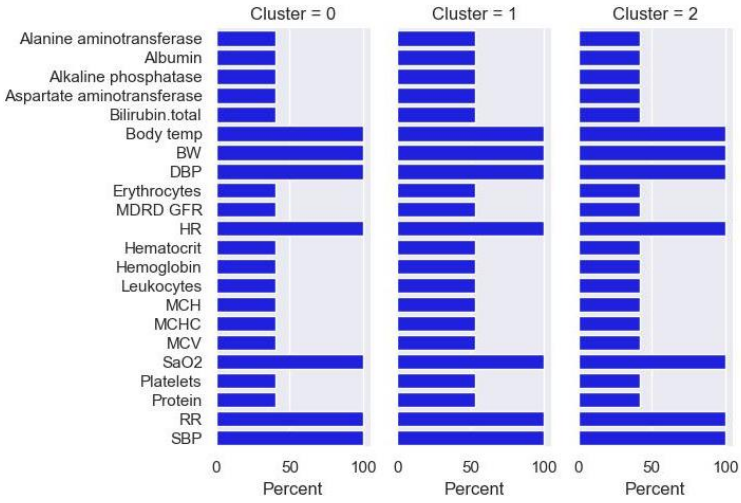

Figure 5 Top: The distribution of demographics within the two clusters and the overall patients in Case1 (0 day of hospital visit only). Those patients labeled as NonAssigned were not considered for clustering due to absence of any clinical data at the time of the visit. Bottom: The bar graph depicts the frequency of each measurement within the clusters.

In Case4, average across 14 days, the differences in the average age increases as the number of clusters increase from two to three and further increases for four clusters, shown in Figure 6, Figure 7 and Figure 8. For the two clusters the average age is 42 years for Cluster1 vs 48 years for Cluster0. Unlike the Case1 (clinical data at 0 day only), the Case4 (average across 14 days) resulted in greater separation in the clinical data for the two clusters (see Figure 4 vs Figure 6). This is also evident from the silhouette method that evaluates the goodness of clustering. The average silhouette coefficient that measures the separation between clusters is 0.84 for Case4 compared to 0.3 for Case1, that is closer to the ideal value of 1. All of the patients in Cluster1 of the two clusters had erythrocytes, hematocrit, hemoglobin, leukocytes, MCH, MCHC, MCV, and platelets reported compared to 40% in Cluster0. The separation or differences in the clinical data reduces when the number of clusters increases from 2 to 4 and 4. Also, evident from the average silhouette coefficient. It decreases from 0.84 to 0.38 and 0.07 for the number of clusters two, four and three respectively.

	Overall	0	1
n	1763	1756	7
Age	48.4 ± 25.0	48.4 ± 25.0	42.1 ± 23.6
Female	909 (51.6%)	905 (51.5%)	4 (57.1%)
White	1473 (83.6%)	1466 (83.5%)	7 (100.0%)
Black	163 (9.2%)	163 (9.3%)	0 (0.0%)
Asian	114 (6.5%)	114 (6.5%)	0 (0.0%)
Other	13 (0.7%)	13 (0.7%)	0 (0.0%)

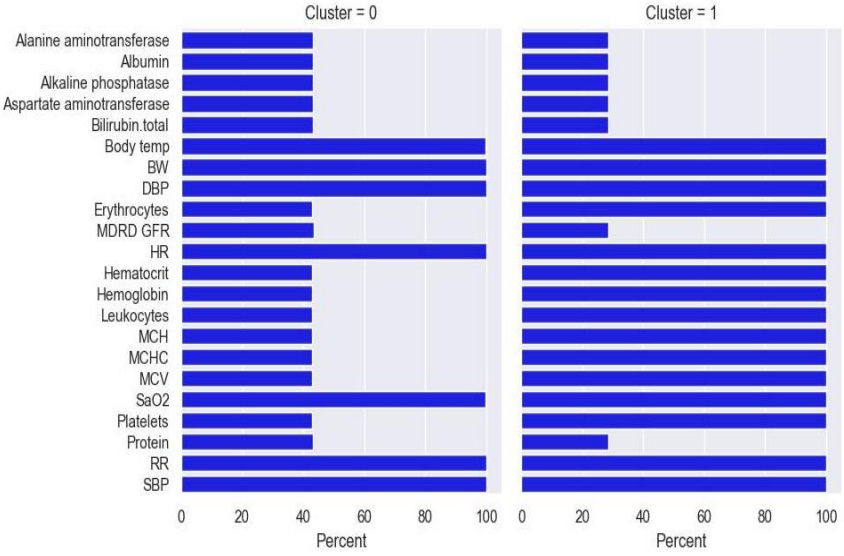

Figure 6 Top: The distribution of demographics within the two clusters and the overall patients in Case4 (average across 14 days). Bottom: The bar graph depicts the frequency of each measurement within the clusters.

	Overall	0	1	2
n	1763	1004	746	13
Age	48.4 ± 25.0	44.6 ± 26.1	53.3 ± 22.2	61.2 ± 29.2
Female	909 (51.6%)	532 (53.0%)	371 (49.7%)	6 (46.2%)
White	1473 (83.6%)	838 (83.5%)	623 (83.5%)	12 (92.3%)
Black	163 (9.2%)	96 (9.6%)	67 (9.0%)	0 (0.0%)
Asian	114 (6.5%)	62 (6.2%)	52 (7.0%)	0 (0.0%)
Other	13 (0.7%)	8 (0.8%)	4 (0.5%)	1 (7.7%)

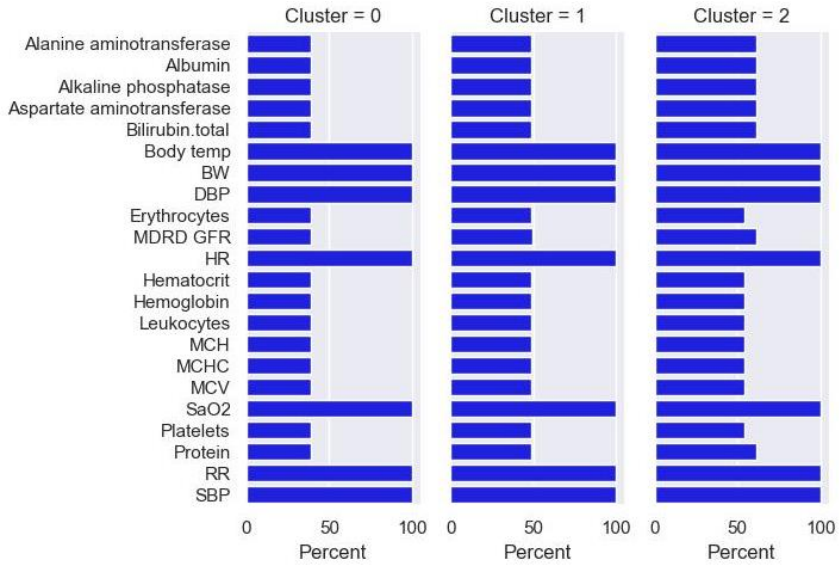

Figure 7 Top: The distribution of demographics within the three clusters and the overall patients in Case4 (average across 14 days). Bottom: The bar graph depicts the frequency of each measurement within the clusters.

	Overall	0	1	2	3
n	1763	1678	7	72	6
Age	48.4 ± 25.0	47.9 ± 24.9	42.1 ± 23.6	58.2 ± 21.8	83.5 ± 16.6
Female	909 (51.6%)	870 (51.8%)	4 (57.1%)	33 (45.8%)	2 (33.3%)
White	1473 (83.6%)	1405 (83.7%)	7 (100.0%)	56 (77.8%)	5 (83.3%)
Black	163 (9.2%)	155 (9.2%)	0 (0.0%)	8 (11.1%)	0 (0.0%)
Asian	114 (6.5%)	108 (6.4%)	0 (0.0%)	6 (8.3%)	0 (0.0%)
Other	13 (0.7%)	10 (0.6%)	0 (0.0%)	2 (2.8%)	1 (16.7%)

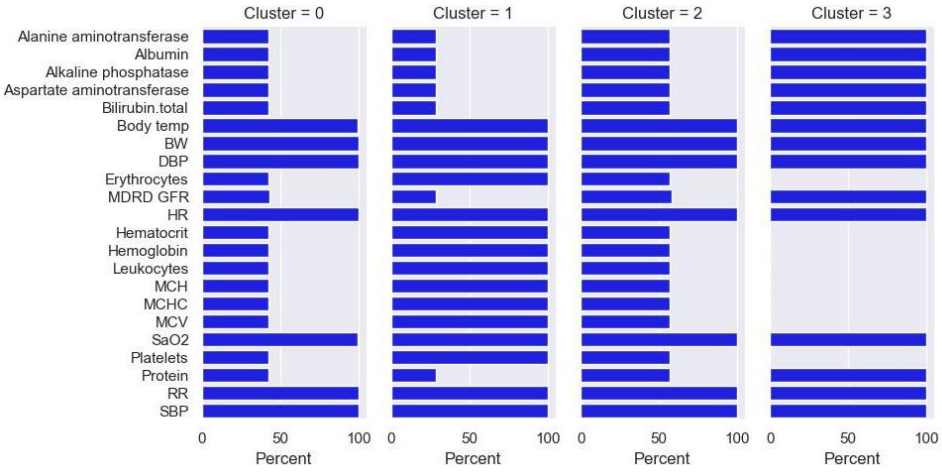

Figure 8 Top: The distribution of demographics within the four clusters and the overall patients in Case4 (average across 14 days). Bottom: The bar graph depicts the frequency of each measurement within the clusters.

Discussion:

COVID19 disease severity ranges from mild to severe with pneumonia, respiratory complications and death. The application of clustering method on the clinical data for 22 types of labs and vitals of 1763 patients identified 2 clusters and best segmentation of phenotypes was identified from averaged clinical data across 14 days. Goodness of cluster was evaluated using the metric average silhouette coefficient, and the best separation between clusters (0.84 silhouette coefficient) was obtained for 2 clusters compared to the grouping into 3 and 4. The Cluster0 is older with higher prevalence of alanine aminotransferase, albumin, alkaline phosphatase, aspartate aminotransferase, bilirubin, MDRD GFR (40%), protein compared to Cluster1 (30%). The Cluster1 has higher frequency of erythrocytes, hematocrit, hemoglobin, leukocytes, MCH, MCHC, MCV and platelets.

References:

[1] Struyf T, Deeks JJ, Dinnes J, et al., "Signs and symptoms to determine if a patient presenting in primary care or hospital outpatient settings has COVID-19 disease.," *Cochrane Database Syst Rev. ,* vol. 7, 2020.

[2] Ito F, Terai H, Kondo M, et al., "Cluster analysis of long COVID in Japan and association of its trajectory of symptoms and quality of life.," *BMJ Open Respir Res.,* vol. 11, no. 1, 2024.

[3] Fernández-de-Las-Peñas C, Martín-Guerrero JD, Florencio LL, et al., "Clustering analysis reveals different profiles associating long-term post-COVID symptoms, COVID-19 symptoms at hospital admission and previous medical co-morbidities in previously hospitalized COVID-19 survivors.," *Infection,* vol. 51, no. 1, pp. 61-69, 2023.

[4] Reese, Justin T.Spratt, HeidiThorpe, Lorna E. et al., "Generalisable long COVID subtypes: findings from the NIH N3C and RECOVER programmes," *eBioMedicine,* vol. 87, p. 104413, 2023.

[5] Pedregosa et al., "Scikit-learn: Machine Learning in Python," vol. 12, no. 85, pp. 2825-2830, 2011.